

基于贝叶斯网络的稀疏出租车 GPS 轨迹路径还原方法

李广耀, 黄正锋*, 楼乐依

(宁波大学 海运学院, 浙江 宁波 315832)

摘要: 为提高出租车 GPS 大数据的可用性, 提出一种基于贝叶斯网络研究稀疏出租车 GPS 轨迹路径还原的方法. 与传统仅基于时空变量的研究方法不同, 新算法同时考虑天气条件、驾驶员特性、车辆行驶特性与出租车的载客状态等因素来进行路径还原预测. 以宁波市体育中心周围的路网为例, 将出租车服务信息平台的 GPS 轨迹数据作为测试对象, 验证本文方法的适用性. 结果显示, 基于多因素的贝叶斯网络方法在还原精度方面(达到 91.4%)优于 Logit 选择模型. 此外, 新算法尤其适用于出租车轨迹数据缺失率较高的场景, 比如缺失轨迹点跨度在 5 min 左右.

关键词: 稀疏出租车 GPS 数据; 贝叶斯网络; 多因素; 轨迹还原; 缺失率

中图分类号: U491.1

文献标志码: A

文章编号: 1001-5132 (2021) 02-0017-08

随着信息通讯手段的进步和车载导航仪、智能手机等设备的普及, 移动出行数据的获取变得更加容易. 以出租车为例, 其移动出行数据(主要是 GPS 数据)蕴含着丰富的车辆状态信息, 如经纬度坐标、车牌号、时间、车辆载客状态、实时车速、路段名称等. 通过对这些数据进行处理, 可以挖掘出用户的行为信息^[1]. 例如, 通过出行者位置与路网的匹配工作, 实现出行者路径轨迹判断, 可以进一步获知出行者路径选择行为特征^[2-4]. 针对密集 GPS 轨迹数据, 通过地图匹配将轨迹点位映射至路网, 可以较为容易地判断行走轨迹. 然而由于定位设备质量、周边环境干扰以及数据传输不确定性等问题, 造成出租车 GPS 数据缺失率较高. 当采样率较低时, 通过点位数据来还原车辆行走轨迹就变得异常困难. 比如一辆 $50 \text{ km} \cdot \text{h}^{-1}$ 的汽车若在 2 min 内没有 GPS 数据更新, 会致使 1.6 km 区间内的车辆行驶轨迹无法直接重现; 如果此区间内有多个交叉口, 则所选择路径就有多种可能, 而如何准确还原其行走路径就是个难题. 从另一层面来看, 若能够有效解决此问题, 则将降低数据采集与存储成本.

针对稀疏轨迹路径还原问题, 常用方法有以

下几种: (1)最短路径法; (2)平均弗雷歇距离法 (Average Fréchet Distance, AFD); (3)隐马尔科夫模型 (Hidden Markov Model, HMM) 相关方法; (4)其他评估优化方法.

最短路径法是较早被用来还原稀疏 GPS 行走路径的方法. 如 Bierlaire 等^[5]在研究路网较稀疏、可能的行走路径数量较少时, 使用最短路径法来还原实际行走轨迹, 其精度较高, 但当面向密集复杂的城市路网却不一定合适. Brakatsoulas 等^[6]提出一种基于曲线相似度的地图匹配算法, 使用 AFD 衡量 GPS 序列和候选路段序列的匹配度, 将匹配度最高的路径作为最终匹配路径, 但应用较为复杂. 现今应用较多的是基于 HMM 的一系列地图匹配方法^[7]. Lou 等^[8]基于 HMM 提出 ST-Matching 算法, 其还原精度与运行时间均优于 AFD 法. 同时在路径还原中还涉及到许多混合型 HMM 方法^[9-10], 但这些方法一般不适用于 GPS 高缺失率的情形. 对于更低频 GPS 采样数据的路径还原问题, 一些研究则采用评估优化相关方法. 王龙飞等^[11]在考虑多种因素的基础上, 利用逼近理想解排序的 TOPSIS 法获取最优轨迹, 但其局限性表现在要求路网为方格形状. Shuidong 等^[12]则提出了一种分

收稿日期: 2020-07-06.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 宁波市自然科学基金(2019A610040); 国家自然科学基金(51408321); 浙江省自然科学基金(LY18E080009).

第一作者: 李广耀(1995-), 男, 山东青岛人, 在读硕士研究生, 主要研究方向: 交通运输规划与管理. E-mail: 1518432988@qq.com

*通信作者: 黄正锋(1986-), 男, 浙江金华人, 博士/副教授, 主要研究方向: 交通运输工程. E-mail: huangzhengfeng@nbu.edu.cn

布式鲁棒优化方法来轨迹还原,但数据存在与现实情况不一致的问题。

以上诸多研究方法都未将区域交通运行特征、驾驶人路径选择行为特征、其他环境因素等纳入考虑,因此,这些方法不一定适用于GPS轨迹缺失率较高的路网场景。本文主要针对数据缺失率较高(无轨迹点位时间达到5 min及以上)的情形,在采集城市出租车GPS数据基础上,综合考虑时间、空间、司机特性、环境特性、营运特性等因素,结合实际路网情况重建路径轨迹。具体来说,本文利用密集GPS数据集结合驾驶员、天气状况等相关属性数值,采用贝叶斯网络模型进行样本训练,从而建立路径还原的贝叶斯网络,利用建立的贝叶斯网络对稀疏GPS轨迹路径进行还原。

1 问题描述

轨迹还原问题可描述如下:给定某个路网区域 $A_{\text{Area}} = \{V, E\}$, 其中, E 为路段集, 代表路网中的路段; V 为顶点集, 代表路网中的路口或转折点。路段信息包括路段编号、路段长度、路段起终点, 表示为 $e_i = (e_{i,\text{id}}, e_{i,l}, e_{i,\text{start}}, e_{i,\text{end}})$, 其中, $e_i \in E$, $e_{i,\text{start}} \in V$, $e_{i,\text{end}} \in V$ 。多个路段首尾相连即为路径, 用字母 R 表示为 $R: e_1 \rightarrow \dots \rightarrow e_i \rightarrow e_{i+1} \rightarrow \dots \rightarrow e_m$, 且 $e_{i,\text{end}} = e_{i+1,\text{start}}$, $1 \leq i < m$ 。一个GPS日志是一系列GPS点的集合, 用 $g = \{p_1, p_2, \dots, p_i, \dots, p_n\}$ 表示, 每个GPS点包含经度、纬度、时间戳、速度和所在路段信息, 即 $p_i = \{p_{i,\text{lng}}, p_{i,\text{lat}}, p_{i,t}, p_{i,v}, p_{i,e}\}$ 。针对GPS日志中跨越多个路段的前后2个点位 p_i 和 p_{i+1} , 在GPS中间轨迹数据丢失的条件下, 如何复原其真实行驶路径是本文研究的问题。

2 贝叶斯网络模型

贝叶斯网络是一种基于概率推理的有向图模型, 其通过可视化网络模型来表达问题领域中变量的依赖和关联关系, 适用于不确定性知识的表达和推理。不确定性知识在贝叶斯网络中被表示成一个由节点集 $U = \{X_1, X_2, \dots, X_n\}$ 、有向弧段和网络参数 θ 组成的有向无环图(Directed Acyclic Graph, DAG), 有向弧段则代表各个变量之间的因果关系(由父节点指向其子节点), 网络参数 θ 表示网络中父节点指向子节点的条件概率集合。

$$\theta = \{\theta_{X_1|p_a(X_1)}, \theta_{X_2|p_a(X_2)}, \dots, \theta_{X_i|p_a(X_i)}, \dots, \theta_{X_n|p_a(X_n)}\}, \quad (1)$$

式中: $p_a(X_i)$ 表示子节点 X_i 的父节点集; $\theta_{X_i|p_a(X_i)}$ 表示 $p_a(X_i)$ 节点对应的条件概率表。

贝叶斯网络模型的构建主要包括特征因素提取、结构学习、求解参数、网络推理四步, 其中结构学习是最关键的步骤。最常见方法是基于评分搜索算法, 其原理是在所有节点的结构空间内, 按照一定搜索策略和评分准则找出最佳的贝叶斯网络结构。应用于贝叶斯网络结构学习的评分函数主要有基于贝叶斯统计的评分函数和基于信息论的评分函数, 前者精度较高; 当样本数据量非常大时, 则可以考虑后者。

基于评分搜索的算法主要有穷举搜索法、K2算法和爬山算法等^[13-19], 其中K2算法应用较为广泛。本文网络节点和样本数据量适中, 因此, 选用贝叶斯评分函数作为结构学习的评分函数, 应用K2算法获得贝叶斯网络结构, 相应贝叶斯评分函数公式具体如下:

$$K_2(G, D) = \sum_{i=1}^n \sum_{j=1}^{q_i} \left(\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right) + \log(P(g)), \quad (2)$$

式中: G 代表网络结构, 且网络中的任何一个节点变量都属于 U ; D 代表样本数据集; q_i 代表节点 X_i 的父节点取值组合; r_i 代表节点 X_i 的状态取值总数; N_{ijk} 代表数据集中的节点 X_i 取值为 k , 并且父节点处于第 j 种取值时组合的数量 $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$; $P(g)$ 代表网络形态 G 的先验概率(假设网络形态发生概率服从均匀分布)。

3 数据来源与处理

稀疏GPS轨迹实际上是一条等时GPS轨迹丢失了一些点位后的轨迹链信息, 其还原工作的难点在于如何还原最大时距相邻轨迹点对之间的路径。本文构造稀疏GPS轨迹点时, 并非将完整的GPS轨迹点进行随机打断操作, 而是特意选定相距较远的2个GPS点对, 将中间GPS点位舍弃, 由此构造出相对较大的时距轨迹点对场景。通过宁波市出租车服务信息管理平台, 特别选取了密集出租车GPS轨迹数据(采样周期为15 s)与驾驶员的

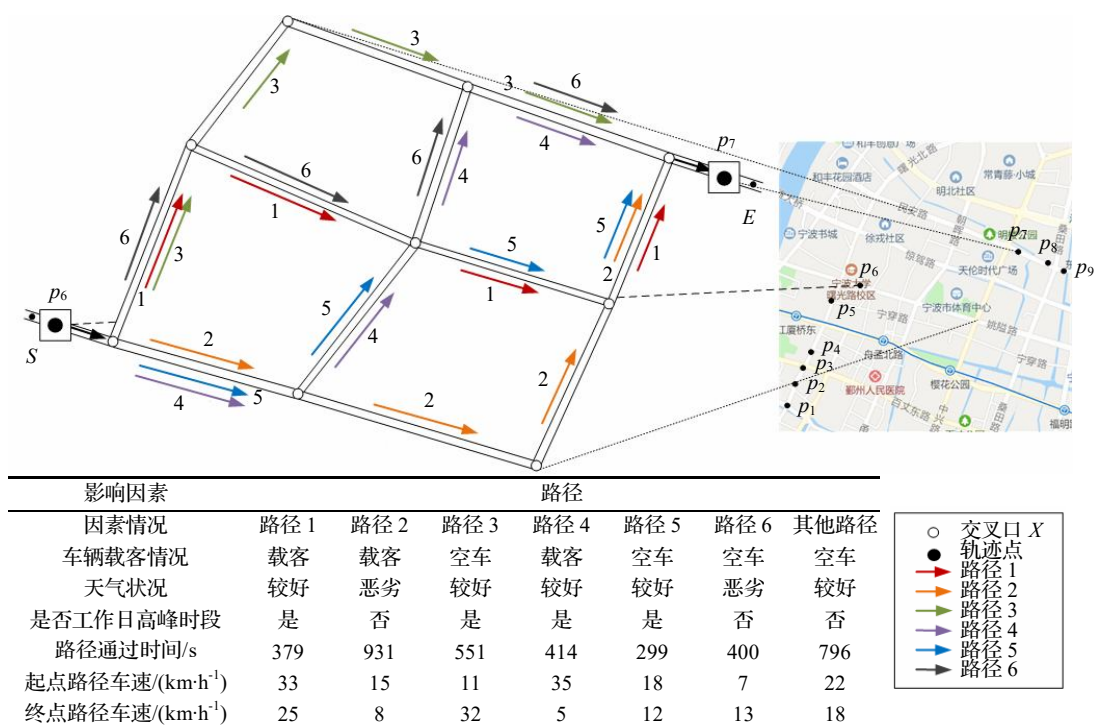


图1 调查范围与轨迹路径示意图

基本信息, 定位好起终路段, 将一部分样本作缺失处理, 保留离路段中心点位置最近的 GPS 点位作为起终点, 并将中途轨迹点作丢弃处理, 形成稀疏 GPS 轨迹, 然后将另一部分未处理的样本作为输入数据, 最后对处理过的稀疏 GPS 轨迹路径进行还原.

本文选定宁波市体育中心周围的路网作为研究范围(图 1), 起终点分别为 p_6 和 p_7 , p_6 和 p_7 点位之间的时间差在 5 min 左右. 通过剔除异常数据, 最终获得有效数据数 15 248 条, 有效率为 90%, 时间分布区间为 2017 年 12 月 9 日~2018 年 12 月 31 日, 日均有效数据 41 条, 应用 ARCGIS 软件将这些数据匹配至路网, 与轨迹路径相关联, 获得 18 种可选路径. 为合理选取有限条路径作为备选集, 根据数据分布规律设置以下原则: 生成最多 7 条备选路径, 当实际轨迹路径超过 7 条时, 将选择概率较小的路径都纳入第 7 条名义路径. 本文在备选路径中一共设置 7 条, 除了最高选择比的 6 条路径, 最后 1 条为剩余路径集合, 具体如图 1 所示.

车辆在通过设定起终点路段时, 鉴于各自 GPS 点位不可能完全重合, 因此需要将上下游路段起终 GPS 点位都映射到一个具有参考性的位置来进行操作, 才有利于在换算通行时间、速度等数

值时, 保证相关输入数据的一致性. 文中使用路段中心点作为统一参考位置, 将上下游路段 GPS 点位都映射到路段中心点. 由于映射过程中需要对起终点时间戳进行修正, 因此具体以图 2 为例对修正方法进行说明.

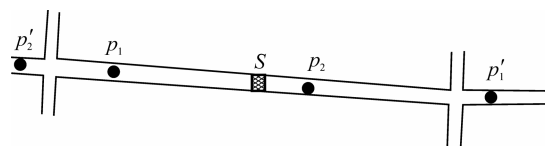


图2 起始路段车辆点位示意图

图 2 中, S 为设置的起始路段中心点, 首先需要找到以 S 为途径点的相邻采集点位, 比如 p_1 、 p_2 为不同车辆经过起始路段时被采集到的点位, 则 p'_1 是 p_1 的后一个有效相邻采样点, p'_2 是 p_2 的前一个有效相邻采样点, 计算车辆 i 在路段中心点 S 处的速度结果如下:

$$v_{is} = l_{i1} / \Delta t, \quad (3)$$

式中: v_{is} 代表车辆 i 经过路段中心点 S 处的车速; l_{i1} 代表车辆 i 在起始路段上有效采样点间的距离; Δt 为采样时间间距.

计算车辆 i 在路段中心点 S 处的时间戳修正结果如下:

$$t_{is} = t'_{is} + \delta l_{i2} / t_{is}, \quad (4)$$

式中: t_{is} 代表车辆 i 在途径路段中心点 S 处的时间戳; t'_{is} 代表车辆 i 在距离路段中心点 S 最近点位的时间戳; l_{i2} 代表最近点与路段中心点间的距离; δ 代表判断变量, 若最近点在路段中心点上游, 则取值为 1, 否则取值为 -1.

4 路径还原影响因素选择

针对稀疏 GPS 轨迹, 总结与路径选择行为关联的各类因素如下:

(1) 营运特性. 车辆的不同载客状态能够影响路径选择结果. 空载车辆一般优先选择途径热点区域的路段, 载客车辆则无此考虑^[14]. 另外, 空载车辆驾驶员出于巡视路边扬招乘客的需求, 其运行车速相对较慢.

(2) 车速特性. 路段与路径间的交通状态存在空间关联性^[15]. 起点路段与终点路段作为出租车选择路径必经路段, 其车速可以不同程度反映各条路径的交通状态. 比如某路段与相应路径的流量关联性较大, 若在离散时段(如 5 min)范围内, 采集途径该路段所有出租车速度平均值, 则该值大小能在很大程度上反映相应路径的出行时长.

(3) 环境特性. 天气状况对路径选择结果有重要影响. 在恶劣天气下, 不确定因素增加, 道路能见度、安全状况有所变化, 驾驶员有可能根据个体偏好及出行经验调整出行路径^[16](本文天气数据来自网络: <http://www.tianqihoubao.com/lishi/>).

(4) 驾驶员特性. 外部环境相同时, 驾驶员的自身特性(如年龄、驾龄、性别等)差异也能够影响路径决策结果. 张卫华等^[20]通过 Logit 建模发现, 年轻、短驾龄特性的驾驶员倾向于灵活择路, 而年长、长驾龄特性的驾驶员则容易坚持常开路径.

营运、车速、环境三类数据与路径选择结果的直观影响关系如图 1 所示, 在选取的 7 个样本中可以发现, 不同路径选择所对应的因素取值有所差异.

本文结合相关文献与专家经验, 经进一步的相关性分析, 筛选得到 10 个节点变量. 以影响司机路径选择的因素作为网络输入节点, 相关变量为车辆载客状态(X_1)、天气状况(X_2)、是否工作日高峰时段(X_3)、通过整条路径时长(X_4)、起点路段车速(X_5)、终点路段车速(X_6)、司机驾龄(X_7)、司机

年龄(X_8)、司机性别(X_9), 并将供司机选择的有效备选路径作为根节点(X_{10}).

5 因素特征提取

特征提取在对因素取值区间进行界定以及对相关连续值进行离散化处理时, 会对轨迹还原模型的准确性有较大影响. 因此, 本文在考虑的相关因素中, 将车辆载客状态(X_1)、天气状况(X_2)、是否工作日高峰时段(X_3)、司机性别(X_9)、备选路径(X_{10})为属性变量, 而通过路径的时长(X_4)、起点路段的车速(X_5)、终点路段的车速(X_6)、司机驾龄(X_7)、司机年龄(X_8)为作数量变量. 为满足贝叶斯网络的建模要求, 将属性变量编码处理为虚拟变量, 同时将部分连续变量编码处理为离散变量. 离散化即将数值按值域划分为不相交的若干个值区间, 每个区间对应一个离散值, 将原始数据更新为离散值. 结合数据分布情况对数据进行离散化, 得到表 1 的分类数据.

由表 1 可见, 所有样本中的各类因素取值及占比都较为合理, 与现实情况没有过大反差. 车辆载客状态方面, 由于调查区域为城市核心区, 重车比为 74%, 因此略高于市区平均值. 部分数据取值会比较低, 比如恶劣天气(雨雪、台风、大雾)、高峰时段、女出租车司机占比. 剩余一些数据分布都较为平均, 比如各类出行时间、起终点路段车速、驾龄、司机年龄.

针对不同路径的样本做归类, 在每条路径选择样本中, 发现各类因素特征都有取值, 比如在路径 2 的样本中, 运行时长分别占比为 12.3%、17.0%、12.8%和 57.9%. 这种现象体现了 GPS 轨迹路径还原问题不能仅用单一因素或少量几个因素进行分析, 应该多类因素共同建模判断.

6 基于路径还原问题的贝叶斯网络模型结果

为校验贝叶斯网络模型精度, 将样本数据依照 3:1 的比例划分为训练集和测试集, 划分结果为 11 436 和 3 812 个样本. 利用专家知识和 K2 算法相融合方法获得贝叶斯网络结构, 以样本数据为基础, 利用各变量与选择路径待关注参量相关性大小进行排序, 确定各变量的输入节点顺序为天气、司机年龄、司机性别、是否工作日高峰时段、终点

表 1 路径选择的影响因素分类及占比统计情况(训练集)

相关特性	变量名	符号	变量取值	类样本数/个	类样本数占比/%	相关特性	变量名	符号	变量取值	类样本数/个	类样本数占比/%
营运特性	车辆载客	X_1	空车	3 015	26	司机特性	司机	X_7	<5	1 559	14
	状态		载客	8 421	74		驾龄/a		$5 \leq X_7 < 10$	3 035	26
环境特性	天气状况	X_2	恶劣天气	760	7				$10 \leq X_7 < 15$	2 700	24
			天气较好	10 676	93				≥ 15	4 142	36
	是否工作日	X_3	工作日高峰	1 629	14	司机年	X_8	X_8	<35	1 004	9
	高峰时段		其他时段	9 807	86	龄/岁			$35 \leq X_8 < 45$	3 403	30
车速特性	路径通过	X_4	<300	2 232	19				$45 \leq X_8 < 55$	6 203	54
	时长/s		$300 \leq X_4 < 400$	3 476	30				≥ 55	826	7
			$400 \leq X_4 < 500$	2 235	20	司机性别	X_9	X_9	男	10 673	93
			≥ 500	3 493	31					女	763
	起点路段	X_5	<10	915	8	目标变量	选择路径	X_{10}	路径 1	662	6
	车速/		$10 \leq X_5 < 20$	3 087	27				路径 2	3 951	35
	(km·h ⁻¹)		$20 \leq X_5 < 30$	5 604	49				路径 3	1 298	11
			≥ 30	1 830	16				路径 4	1 285	11
	终点路段	X_6	<10	801	7				路径 5	3 005	26
	车速/		$10 \leq X_6 < 20$	3 774	33				路径 6	606	5
	(km·h ⁻¹)		$20 \leq X_6 < 30$	5 717	50				其他路径	629	6
			≥ 30	1 144	10						

注: 工作日高峰时段为 7:00~8:30 以及 16:30~18:30.

路段车速、司机驾龄、车辆状态、起点路段车速、路径通过时长、选择路径, 设置网络中任一节点最大父节点数量不超过 4, 应用 K2 算法在 Matlab 软件中的 BNT 工具箱完成程序编程, 进行稀疏数据车辆轨迹还原贝叶斯网络结构学习, 以训练集作为数据源, 最终得到贝叶斯网络结构(图 3).

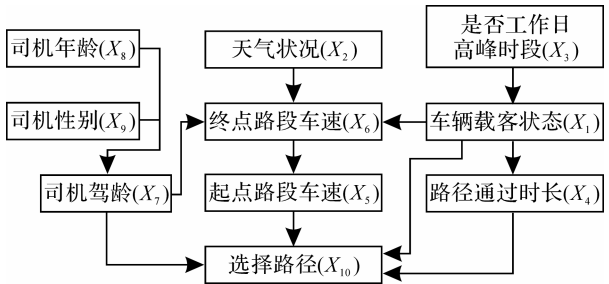


图 3 贝叶斯网络构建结果

根据图 3 可以直观地看到各个节点之间的变量关系, 其中终点路段车速是起点路段车速的父节点, 可以从两个角度理解这个结果. 由于案例研究范围的起终点距离不是特别远, 起终点路段交通流存在一定联系, 因此车速也具有一定的关联性. 起点路段车速与终点路段车速相比, 它与路径

选择的相关性更大, 成为了子节点, 其原因可能是驾驶员在起点路段时, 根据当下路段交通状况对前方各条路径有预判, 对路径选择具有一定指导性; 而终点路段车速无此特征, 从而起点路段的交通状况与车辆选择路径的相关性更大.

7 模型结果检验

7.1 精度验证

利用贝叶斯网络在 Matlab 软件中的 BNT 工具箱来完成程序编程, 进行贝叶斯网络参数学习, 获得对应的条件概率表集合, 而后根据网络结构与对应的条件概率表对测试集的样本进行选择路径预测, 并根据预测结果计算查全率(模型对特定目标的识别能力)和查准率(模型区分特定样本与其他样本的能力), 最终得到驾驶员路径选择识别混淆矩阵(表 2).

由表 2 数据可见, 测试集预测驾驶员选择路径的准确度(模型预测结果的精度)达到 91.4%, 查全率与查准率均高于 86.7%, 其中其他路径样本的查全率和查准率都超过了 93.2%. 由此可得, 该模型

表2 贝叶斯网络模型与多项 Logit 回归模型预测还原路径精度对比表

项目	贝叶斯网络驾驶员路径选择识别混淆矩阵								查全率/%	多项 Logit 模型驾驶员路径选择识别混淆矩阵								查全率/%
	路径	路径	路径	路径	路径	路径	其他路径	路径		路径	路径	路径	路径	路径	其他路径			
	1	2	3	4	5	6		1		2	3	4	5	6				
真实的 选择 路径	路径 1	192	1	1	2	1	0	0	97.5	82	98	4	13	0	0	0	41.6	
	路径 2	6	1 211	22	30	70	6	3	89.8	32	1 215	14	52	1	10	24	90.1	
	路径 3	3	5	389	5	3	2	2	95.1	33	239	46	79	3	8	1	11.2	
	路径 4	4	13	7	384	4	2	3	92.1	22	134	6	209	30	15	1	50.1	
	路径 5	9	60	13	20	922	3	2	89.6	10	14	0	44	916	43	2	89.0	
	路径 6	2	10	0	2	4	209	3	90.9	1	24	4	2	78	120	1	52.2	
	其他路径	0	0	0	0	2	1	179	98.4	1	2	3	3	1	1	171	94.0	
查准率/%	88.9	93.2	90.0	86.7	91.7	93.7	93.2	91.4	45.3	70.4	59.7	52.0	89.0	60.9	85.5	71.7		

预测精度较高, 较为合理。

为验证本文建立的路径还原贝叶斯网络模型在轨迹数据缺失较为严重时, 其还原轨迹路径的相对优越性, 采用与多项 Logit 回归模型进行比较的方法。先对问题进行分析建模, 然后应用 SPSS 统计软件使用最大似然估计法对参数进行标定, 再进行多项 Logit 回归分析, 最终得到模型的预测结果(表 2)。从表 2 数据可见, 多项 Logit 回归的总体预测准确率为 71.7%, 相比贝叶斯网络模型差距明显, 查全率与查准率最小值仅为 11.2% 和 45.3%, 其他路径样本的查全率和查准率为 94.0% 和 85.5%, 略低于贝叶斯网络模型。从模型预测精度矩阵分布上看, 贝叶斯网络模型在各类选择上的预测能力远优于多项 Logit 回归模型, 在用于不同样本时, 其预测能力更为稳定。由此可见, 在稀疏轨迹数据的路径还原领域方面, 贝叶斯网络建模方式的预测准确性更高。

7.2 ROC 曲线分析

利用 SPSS 软件计算出各个阈值下的假阳性率与真阳性率, 真阳性率表示被模型识别为指定路径样本数占实际为该路径样本数的比例, 假阳性率表示被模型错误识别为指定路径样本数占实际非该路径样本数的比例。分别对 7 条路径绘制出 7 条 ROC 曲线, 最后对这 7 条 ROC 曲线取平均, 得到总 ROC 曲线(图 4)。经计算, 贝叶斯网络模型的 AUC 值(曲线下方面积, Area Under Curve)为 0.875, 根据 ROC 曲线评价标准可知, 贝叶斯网络预测模型预测精度为良好。由此可见, 应用贝叶斯网络模

型来还原轨迹路径的方法具有优势。

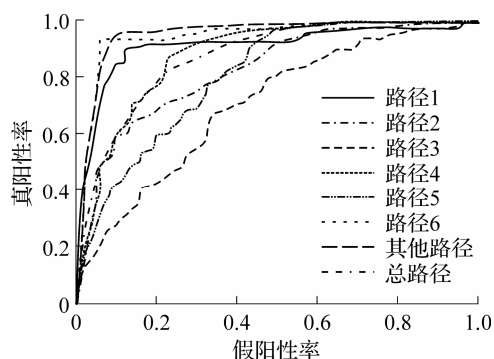


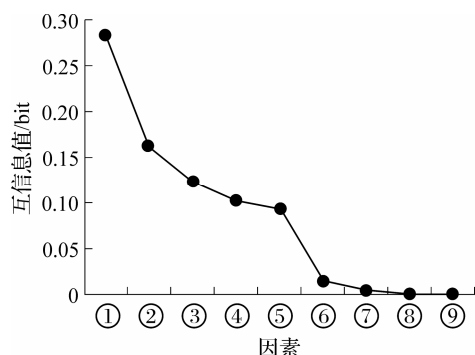
图4 贝叶斯网络的 ROC 曲线分析

从图 4 的 ROC 曲线还可以看出, 其他路径的 ROC 曲线最为贴近左上角, 说明模型对其他路径的预测精度最高。原因在于其涵盖的路径均存在一定绕行情况, 通过时长显著高于占比较大的其他 6 条路径, 区分较为明显。其次则为路径 1 和路径 6 的 ROC 曲线。通过对统计数据进行分析, 选择路径 6 的司机驾龄绝大部分在 15 a 及以上, 推测经验丰富的驾驶员倾向将该路径作为避开拥堵路段的一大选择, 而选择路径 1 的司机驾龄大多数集中在 10 a 以上和 5 a 以下, 这些取值的聚集特征使得路径 1 和 6 的预测精度较高。

7.3 敏感性分析

利用互信息值检验贝叶斯网络的推理结果, 计算出各因素与选择路径的互信息值, 以观察各因素对选择路径的影响程度, 如果互信息值越大, 则表明影响程度越大。由图 5 可见, 对路径选择影响程度较大的因素主要有时长(0.28 bit)、司机驾龄

(0.16 bit)、车辆载客状态(0.12 bit)以及起点路段车速(0.10 bit)。



①时长; ②驾龄; ③车辆载客状态; ④起点车速; ⑤司机年龄;
⑥终点车速; ⑦司机性别; ⑧是否工作日高峰时段; ⑨天气。

图5 各因素对选择路径的互信息值影响度

7.4 轨迹数据缺失时间对模型预测精确度的影响

为了进一步研究轨迹数据缺失时间变化对模型还原精度的影响, 则将研究范围逐步扩大, 如图6所示, 将研究范围由区域1逐步扩展至区域2、区域3、区域4, 终点由 p_2 逐步扩展至 p_3 、 p_4 、 p_5 , 在分别抹去2个点位之间的轨迹点位信息, 数据缺失时间分别为8、12、15 min左右, 参照上文步骤绘制ROC曲线(图7), 最终得到ROC曲线的AUC值

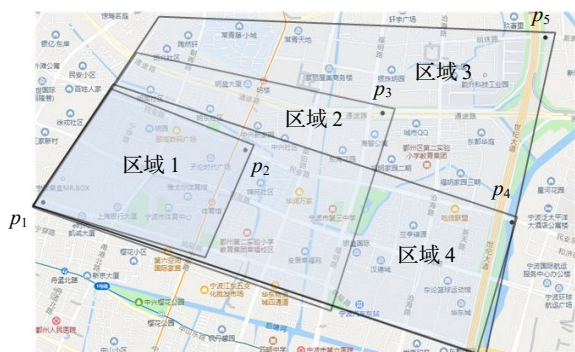


图6 不同数据缺失时间研究范围示意图

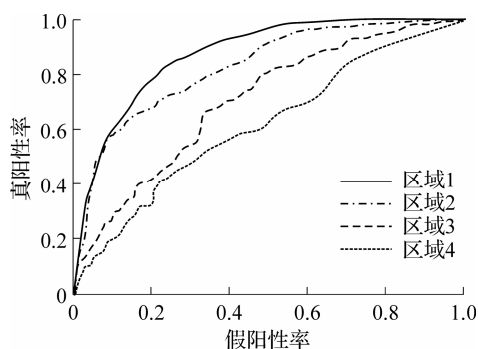


图7 不同缺失时间数据下模型 ROC 曲线

分别是 0.825、0.787 和 0.560。可见, AUC 值随着研究范围的逐渐扩大而逐渐降低, 当数据缺失时间达到 15 min 时, 即研究范围扩展到区域 4 时, 由于选择路径的不确定性增强, 模型预测精度下降, 此时的评价为一般, 甚至接近于差。

8 结论

提出了一种基于稀疏出租车 GPS 轨迹数据的路径还原方法, 综合考虑了包含驾驶员特性、天气与车辆状态等多种因素, 结合宁波市体育中心周围路网的出租车数据对其进行检验。结果表明, 这些因素对于轨迹路径还原精度具有较大影响, 新模型的精度优于多项 Logit 回归模型。

(1)通过K2算法从训练集数据中学习得到路径还原的贝叶斯网络结构, 再对预测集数据进行预测检验, 91.4%以上的准确率表明司机特性等因素与驾驶员选择路径高度相关。

(2)通过对不同缺失率样本数据情况下得到的模型 ROC 曲线图分析, 证实在本文的研究案例中, 该方法适用于 5 min 以上的数据缺失时间, 但是当数据缺失时间达到 15 min 时, 本方法就不具有明显优势。

参考文献:

- [1] Yuan N J, Zheng Y, Xie X, et al. Discovering urban functional zones using latent activity trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3):712-725.
- [2] Chen M, Liu Y, Yu X. Predicting next locations with object clustering and trajectory clustering[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer International Publishing, Singapore, 2015.
- [3] Mori U, Mendiburu A, Álvarez M, et al. A review of travel time estimation and forecasting for advanced traveller information systems[J]. Transportmetrica, 2015, 11(2):119-157.
- [4] Lü L, Chen M, Liu Y, et al. A plane moving average algorithm for short-term traffic flow prediction[C]//Advances in Knowledge Discovery and Data Mining. Springer International Publishing, Singapore, 2015.
- [5] Bierlaire M, Chen J, Newman J. A probabilistic map matching method for smartphone GPS data[J].

- Transportation Research Part C: Emerging Technologies, 2013, 26(1):78-98.
- [6] Brakatsoulas S, Pfoser D, Salas R, et al. On map-matching vehicle tracking data[C]//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway, 2005.
- [7] 高文超, 李国良, 塔娜. 路网匹配算法综述[J]. 软件学报, 2018, 29(2):225-250.
- [8] Lou Y, Zhang C Y, Zheng Y, et al. Map-matching for low-sampling-rate GPS trajectories[EB/OL]. [2020-03-05]. <https://www.researchgate.net/publication/221589740>.
- [9] Ozdemir E, Topcu A E, Ozdemir M K. A hybrid HMM model for travel path inference with sparse GPS samples[J]. Transportation, 2016, 45:233-246.
- [10] Taguchi S, Koide S, Yoshimura T. Online map matching with route prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1):338-347.
- [11] 王龙飞, 陈红, 李杨, 等. 车辆出行轨迹调查分析中的丢点轨迹还原[J]. 计算机应用研究, 2014, 31(1):162-165.
- [12] Shuaidong Z, Kuilin Z. A distributionally robust optimization approach to reconstructing missing locations and paths using high-frequency trajectory data[J]. Transportation Research Part C: Emerging Technologies, 2019, 102:316-335.
- [13] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning, 1992, 9(4):309-347.
- [14] 肖光年, 隗志才, 张春勤. 基于贝叶斯网络和 GPS 轨迹数据的出行方式识别[J]. 统计与决策, 2017(6):75-79.
- [15] Marcot B G, Steventon J D, Sutherland G D, et al. Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation[J]. Canadian Journal of Forest Research, 2006, 36(12):3063-3074.
- [16] Pearl J. Fusion, propagation, and structuring in belief networks[J]. Artificial Intelligence, 1986, 29(3):241-288.
- [17] 韩勇, 樊顺, 周林, 等. 基于聚类算法的出租载客点时空分布特征研究[J]. 中国海洋大学学报(自然科学版), 2019, 49(S1):155-162.
- [18] 刘康, 仇培元, 刘希亮, 等. 利用词向量模型分析城市道路交通空间相关性[J]. 测绘学报, 2017, 46(12):2032-2040.
- [19] 赵晓华, 任贵超, 陈晨, 等. 不良天气下驾驶行为研究综述[J]. 交通信息与安全, 2017, 35(5):70-75; 98.
- [20] 张卫华, 李梦凡. 不同交通信息诱导下驾驶员路径选择行为研究[J]. 重庆交通大学学报(自然科学版), 2018, 37(10):86-93.

Bayesian network-based GPS path restoration for sparse taxi trajectories

LI Guangyao, HUANG Zhengfeng*, LOU Leyi

(Faculty of Maritime and Transportation, Ningbo University, Ningbo 315832, China)

Abstract: In order to improve the availability of taxi GPS big data, a method based on Bayesian network for sparse taxi GPS path restoration is proposed. Different from the traditional research that is only based on spatiotemporal variables, the algorithm takes into account the weather conditions, driver characteristics, vehicle driving characteristics and taxi load status to calculate path restoration prediction. The applicability of the presented method is verified by taking the road network around Ningbo sports center as an example combined with GPS trajectory data collected from the taxi service information management platform for the testing purposes. The case study results show that the Bayesian network method based on multi factors is superior to the Logit selection model in restoration accuracy (up to 91.4%). In addition, the algorithm is especially suitable for the situation with high missing rate of taxi track data, such as the track points of about 5-minute missing span.

Key words: sparse taxi GPS data; Bayesian network; multiple factors; trajectory restoration; missing rate

(责任编辑 章践立)